

## **SOURCE AND ACCURACY STATEMENT FOR THE SURVEY OF PROGRAM DYNAMICS SECOND LONGITUDINAL FILE**

### **DATA COLLECTION AND ESTIMATION**

#### **Source of Data**

The Survey of Program Dynamics (SPD) universe is the noninstitutionalized resident population living in the United States. This population includes people (including children) living in group quarters, such as dormitories, rooming houses, and religious group dwellings. Crew members of merchant vessels, Armed Forces personnel living in military barracks, and institutionalized people, such as correctional facility inmates and nursing home residents, were not eligible to be in the survey. In addition, United States citizens residing abroad were not eligible to be in the survey. Foreign visitors who work or attend school in this country and their families were eligible; all others were not eligible to be in the survey. With the exceptions noted above, people who were at least 15 years of age at the time of the interview were eligible to be asked questions regarding income and job experience.

The calendar year data for 1996 were collected during April, May, and June of 1997 as part of the SPD Bridge Survey. Likewise, the calendar year data for 1997 to 1999 were collected during May, June, and July of 1998 to 2000 as part of the SPD 1998 to 2000 Surveys, respectively. The SPD Bridge calendar and SPD 1998 calendar year files consist principally of the calendar year data for 1996 and 1997, respectively. The SPD second longitudinal file longitudinally combines the data from the SIPP Panels 1992 and 1993, and the SPD Bridge file and SPD 1998 to 2000 calendar year files.

The goal of the SPD program is to provide policy makers a survey to assess the effects of the recent welfare reforms and how these reforms interact with each other, and with employment, income, and family circumstances. The SPD program eventually spans from the pre-reform through the post-reform period, 1992-2002. In order to obtain information about past economic history, employment, income, and program participation, two retired SIPP panels 1992 and 1993 were chosen as the SPD sample. A full potential of the SPD data is generally achieved when using the SPD second longitudinal file until the release of other subsequent SPD longitudinal files.

The SPD Bridge Survey data was collected in 1997 and intended to be a connection run between the SIPP and SPD data. Data that was merged from the previous SIPP surveys, the SPD Bridge survey, and the subsequent SPD survey (for example, SPD 1998) should give us the necessary pre-reform and post-reform information for sampled households.

## Background of SIPP 1992 and 1993 Panels and SPD Bridge Survey

The 1992 and 1993 SIPP panel samples were located in 284 Primary Sampling Units (PSUs), each consisting of a county or a group of contiguous counties. Within these PSUs, expected clusters of two or four living quarters (LQs) were systematically selected from lists of addresses prepared for the 1980 decennial census to form the bulk of the sample. To account for LQs built within each of the sample areas after the 1980 census, a sample was drawn of permits issued for construction of residential LQs up until shortly before the beginning of the panel. In jurisdictions that do not issue building permits, small land areas were sampled and the LQs within were listed by field personnel and then sub-sampled. In addition, sample LQs were selected from supplemental frames that included LQs identified as missed in the 1980 census and group quarters (GQs).

At the time of the initial visit of the SIPP panels, the occupants of about 19,600 living quarters were interviewed for the 1992 panel and 19,900 were interviewed for the 1993 panel. This accounts for approximately 72% (1992) and 73% (1993) of the LQs originally designated for the SIPP samples. Approximately 21% (1992) and 20% (1993) of the designated LQs were found to be vacant, demolished, converted to nonresidential use, or otherwise ineligible for the survey. The remainder, approximately 2000 LQs, were not interviewed because the occupants refused to be interviewed, could not be found at home, were temporarily absent, or otherwise unavailable. Thus, occupants of about 91% of all eligible LQs participated in the first interview of the 1992 and 1993 SIPP panels.

For the remaining nine interviews, only original sample people (those in Wave 1 sample households and interviewed in Wave 1) and people living with them were eligible to be interviewed. With certain restrictions, original sample people were to be followed even if they moved to a new address. When original sample people moved without leaving a forwarding address or moved to extremely remote parts of the country and no telephone number was available, additional non-interviews resulted.

The 1992 10-Wave Longitudinal File consists of data collected from February 1992 to April 1995. Data for up to 39 reference months are available for people on this file. The 1993 Nine-Wave Longitudinal File consists of data collected from February 1993 to January 1996. Data for up to 36 reference months are available for people on this file.

Tables 1a-1c indicate the interview and reference months for the collection of data from the 1992 Ten-Wave Longitudinal File, 1993 Nine-Wave Longitudinal File, and the SPD 1998 to 2000 Files. For the SIPP, a person was classified as interviewed or non-interviewed based on the following definitions. (Note: A person may be classified differently for calculating different weights). Interviewed sample people (including children) were defined to be: *those for whom self, proxy, or imputed responses were obtained for each month of the appropriated longitudinal period.*

The months for which people were deceased or residing in an ineligible address were identified on the file. Non-interviewed people were defined to be those for whom neither self nor proxy responses were obtained for one or more months of the appropriate longitudinal period (excluding imputed people and people who died or moved to an ineligible address).

It is estimated that roughly 56,300 (1992) and 57,200 (1993) people were initially designated in the sample for the SIPP. Approximately 51,100 (1992) and 51,900 (1993) people were interviewed in Wave 1; while the balance, residing in the 4,000 (1992 and 1993 combined) living quarters not interviewed at Wave 1 remained anonymous and became the initial source of the person non-response in the weighting procedures. For panel weighting, the eligible sample is considered to be all people initially classified as interviewed with a person non-response rate of 25 percent (1992) and 24 percent (1993). The longitudinal file contains approximately 59,700 (1992) and 62,700 (1993) people in all. This includes the Wave 1 interviewed people and about 8,600 (1992) and 10,600 (1993) people who entered survey households during the panel through births, marriages, and other reasons. Some respondents did not respond to some of the questions; therefore, item non-response rates, especially for sensitive income and money related items, are higher than the person non-response rates given above.

We define the SPD Bridge sample cohort as people in the 1992 and 1993 SIPP panels that were in an interviewed household in the last wave. However, only people considered interviewed (self or proxy or imputed response) longitudinally in SIPP *and* considered interviewed in SPD Bridge were eligible to go on further for the SPD 1998 Survey, since the SPD 1998 Survey was carried out only as a part of a long-term *longitudinal* data collection effort for the SPD. In addition, due to budget constraint, the SPD 1998 Survey was also subject to a sample cut based on the sub-sampling procedure described in the section below.

### **1998 SPD Sub-sampling for SPD 1998 Survey**

Due to budget constraints, the SPD 1998 Survey did not visit all 35,000 Bridge households. The budget only allowed for SPD to visit 21,000 households. Roughly 19,100 cases were sampled in this operation, since we needed to account for an expected 12.5 percent non-response and a growth of 10 percent of the total sample size due to household spawning.

In the sub-sampling (sample cut), the SPD Bridge sample households were demographically divided into six strata as shown in the table at the end of this section. The stratification was performed using the household information collected from the SPD Bridge. In each stratum, the households were sampled independently with the sampling rate as provided in the table below. As indicated among sampling rates in this table, the low income sample households were generally not subjected to the sample cut at all.

As a result of the sample cut, the actual number of the households selected for interview was 19,288. Among the 19,288 households selected for interview and their spawned households, 16,395 households were interviewed.

Strata	Description	Sampling Rate	Designated Number	Projected Interviews
1	Households where the primary family or the primary individual has a total family income below 150% of the poverty threshold	1-in-1	6,182	5,950
2	Households where the primary family or the primary individual has a total family income between 150% and 200% of the poverty threshold and there are children under 18	1-in-1	1,075	1,035
3	Households where the primary family or the primary individual has a total family income above 200% of the poverty threshold and there are children under 18	1-in-1.11	6,623	6,375
4	Households where the primary family or the primary individual has a total family income between 150% and 200% of the poverty threshold and there are no children under 18	1-in-1.22	1,461	1,406
5	Households in the balance	1-in-3.70	3,707	3,568
6	Households entirely institutionalized (Outcome code = 228)	1-in-3.70	81	DK
Total			19,129	18,334

### SPD 1999 Survey

The sample for the SPD 1999 survey is simply a direct continuation of the SPD 1998 sample. Namely, there is no additional sample cut or no bringing back of any addition sample households from either the SPD Bridge non-respondents or the SIPP Panel 1992/1993 non-respondents. In this survey, there are 19,551 sample households (including the spawned households) eligible for interview, and 16,654 sample households were interviewed.

### SPD 2000 Survey

The sample for the SPD 2000 survey consists of two parts. Part 1 of the SPD 2000 sample is simply a direct continuation of the SPD 1999 sample which will be hereinafter referred to as *the SPD core part*. Part 2 is composed of a sub-sample of the SPD Bridge Type A (direct and indirect refusal) and Type D (unlocated movers) non-interviewed households which will be hereinafter referred to as *the SPD Bridge non-interview part*.

The SPD Bridge non-interview part of the sample was selected by classifying the Type A and Type D SPD Bridge non-interviewed households into six strata in the same manner as those prescribed for the 1998 SPD sample cut as shown in the table below. The non-interviewed households in each stratum were then randomly selected at the sample rate specified in the table

below (which is the same sampling rate for the 1998 SPD sample cut.)

Strata	Description <sup>1</sup>	Number of Households	Percent	Sample Selection Factor	Subsample Size
1	Households where the primary family or the primary individual has a total family income below 150% of the poverty threshold	1,444	22.5	1.00	1,444
2	Households where the primary family or the primary individual has a total family income between 150% and 200% of the poverty threshold and there are children under 18	156	2.4	1.00	156
3	Households where the primary family or the primary individual has a total family income above 200% of the poverty threshold and there are children under 18	736	11.5	1.11	663
4	Households where the primary family or the primary individual has a total family income between 150% and 200% of the poverty threshold and there are no children under 18	476	7.4	1.22	390
5	Households in the balance	3,605	56.2	3.70	974
6	Households institutionalized (outcome code = 228)	0	0.0	3.70	0
Total		6,417	100.0		3,627

<sup>1</sup> The characteristics of these households were determined based on Reference Month 4 of Wave 1 of either the SIPP Panel 1992 or 1993, depending on which panel they were originated from.

In the SPD core part of the sample, there are 19,890 sample households (including the spawned households) eligible for interview, and 16,845 sample households were interviewed. In the SPD Bridge non-interview part of the sample, there are 3,584 sample households (including the spawned households) eligible for interview, and 1,871 sample households were interviewed.

*However, the SPD second longitudinal file includes only the SPD core part of the sample. The SPD Bridge non-interview part of the sample will be included in the SPD third longitudinal file. Therefore, the current source and accuracy statement is prepared with the consideration that the SPD Bridge non-interview part of the sample is not included on the SPD second longitudinal file.*

## ESTIMATION

For estimation of the characteristics of the population in the SPD universe, six weights are provided on the SPD second longitudinal file as listed below.

1. Traditional longitudinal panel weight (SPDLNWGT)
2. 1999 quasi-longitudinal panel weight (SPDQLW99)
3. 2000 quasi-longitudinal panel weight (SPDQLW00)
4. Longitudinal annual weight (SPDAWTRD) associated with the traditional longitudinal panel weight (SPDLNWGT)
5. Longitudinal annual weight (SPDAWQ99) associated with the traditional longitudinal panel weight (SPDQLW99)
6. Longitudinal annual weight (SPDAWQ00) associated with the traditional longitudinal panel weight (SPDQLW00)

The procedure for calculating each of the above six weights is described earlier in the Weighting section.

For the traditional longitudinal panel weight, all sample persons classified as a traditional longitudinal panel interviewed person were assigned a positive final traditional longitudinal panel weight ( $SPDLNWGT > 0$ ), while all those classified otherwise were assigned a zero final traditional longitudinal panel weight ( $SPDLNWGT = 0$ .) The traditional longitudinal panel interviewed persons are those who were classified as interviewed (self, proxy, or imputed) in the entire SIPP (all the reference months) and the SPD Bridge, then survived the 1998 sample cut, and were classified as interviewed in the SPD 1998 to SPD 2000.

For the 1999 quasi-longitudinal panel weight, all sample persons classified as a 1999 quasi-longitudinal panel interviewed person were assigned a positive final 1999 quasi-longitudinal panel weight ( $SPDQLW99 > 0$ ), while all those classified otherwise were assigned a zero final 1999 quasi-longitudinal panel weight ( $SPDQLW99 = 0$ .) The 1999 quasi-longitudinal panel interviewed persons are those who were classified as interviewed (self, proxy, or imputed) in the entire SIPP (all the reference months) and the SPD Bridge, then survived the 1998 sample cut, and were classified as interviewed in the SPD 1999 but were or were not classified as interviewed in the SPD 1998 and/or SPD 2000.

For the 2000 quasi-longitudinal panel weight, all sample persons classified as a 2000 quasi-longitudinal panel interviewed person were assigned a positive final 2000 quasi-longitudinal panel weight ( $SPDQLW00 > 0$ ), while all those classified otherwise were assigned a zero final 2000 quasi-longitudinal panel weight ( $SPDQLW00 = 0$ .) The 2000 quasi-longitudinal panel interviewed persons are those who were classified as interviewed (self, proxy, or imputed) in the entire SIPP (all the reference months) and the SPD Bridge, then survived the 1998 sample cut, and were classified as interviewed in the SPD 2000 but were or were not classified as interviewed in the SPD 1998 and/or SPD 1999.

For the longitudinal annual weight (SPDAWTRD) associated with the traditional longitudinal panel weight (SPDLNWGT), only the following sample people were assigned a positive SPDAWTRD: the original sample persons who were classified as interviewed (self, proxy, or imputed) in the entire SIPP (all the reference months) and the SPD Bridge, then survived the 1998 sample cut, and were classified as interviewed in the SPD 1998 to SPD 2000; and the non-original (200+ level) sample children born after Wave 1 of the SIPP Panel 1992 or 1993 whose designated parents in 2000 have a positive SPDLNWGT. All other sample persons were assigned a zero SPDAWTRD.

For the longitudinal annual weight (SPDAWQ99) associated with the 1999 quasi-longitudinal panel weight (SPDQLW99), only the following sample people were assigned a positive SPDAWQ99: the original sample persons who were classified as interviewed (self, proxy, or imputed) in the entire SIPP (all the reference months) and the SPD Bridge, then survived the 1998 sample cut, and were classified as interviewed in the SPD 1999 but were or were not classified as interviewed in the SPD 1998 and/or SPD 2000; and the non-original (200+ level) sample children born after Wave 1 of the SIPP Panel 1992 or 1993 whose designated parents in 2000 have a positive SPDQLW99. All other sample persons were assigned a zero SPDAWQ99.

For the longitudinal annual weight (SPDAWQ00) associated with the 2000 quasi-longitudinal panel weight (SPDQLW00), only the following sample people were assigned a positive SPDAWQ00: the original sample persons who were classified as interviewed (self, proxy, or imputed) in the entire SIPP (all the reference months) and the SPD Bridge, then survived the 1998 sample cut, and were classified as interviewed in the SPD 2000 but were or were not classified as interviewed in the SPD 1998 and/or SPD 1999; and the non-original (200+ level) sample children born after Wave 1 of the SIPP Panel 1992 or 1993 whose designated parents in 2000 have a positive SPDQLW00. All other sample persons were assigned a zero SPDAWQ00.

The traditional longitudinal panel weight (SPDLNWGT), 1999 quasi-longitudinal panel weight (SPDQLW99), and 2000 quasi-longitudinal panel weight (SPDQLW00) are provided for the estimation of characteristics of the population of the SPD panel universe over the whole or part of the time span between 1992 to 2000. The advantage of using SPDLNWGT for analyses and estimates is that all the original sample people with positive SPDLNWGT have the more data across time from 1992 to 2000 than other cohorts of original sample people. However, any analyses and estimates using SPDLNWGT will automatically be limited to using only the data collected from the original sample people who were interviewed in each and every wave of either the SIPP Panel 1992 or 1993, and in the SPD 1997 through 2000. Therefore, this limitation precludes us from taking the advantage of the data collected from the original sample people who were interviewed in every wave in the SIPP Panel 1992 or 1993, and in the SPD Bridge, survived the SPD 1998 sample cut (reference 2) *but were not interviewed in all of the SPD 1998, 1999, or 2000*. The data from these original sample people is particularly important for evaluating the effectiveness of the money incentive having been offered to Type A (interview refusal) or potential Type A sample households in the SPD 1999 and 2000. Consequently, the 1999 quasi-longitudinal panel weight (SPDQLW99) and the 2000 quasi-longitudinal panel weight (SPDQLW00) are derived to take advantage of the data from these original sample people.

On the SPD second longitudinal file (or any other SPD longitudinal files), 1992 calendar year data is only available for the sample persons that originate from the SIPP Panel 1992. Therefore, for any 1992 estimates created using any one of the traditional longitudinal panel weight (SPDLNWGT), 1999 quasi-longitudinal panel weight (SPDQLW99), and 2000 quasi-longitudinal panel weight (SPDQLW00), the effective panel weight can be approximately taken as two times the panel weight (SPDLNWGT or SPDQLW99 or SPDQLW00) provided on the SPD second longitudinal file. The factor of two is consistent with the combined panel factor of 0.5 used in the weighting as described in the Weighting section. The variances of the 1992 estimates must be consistently approximated by inflating them by two as well.

When calculating annual estimates using any one of the three longitudinal annual weights (SPDAWTRD, SPDAWQ99, and SPDAWQ00), include only the weighted non-original sample children born during or in prior calendar years. The non-original sample children weights, for those aged eight or less, are intended only to approximate growth in the universe of the non-institutionalized residents in the United States since 1993. Some caution should be given when interpreting results for children aged 8 or less. The annual estimates of the characteristics of these children for a calendar year closer to 2000 are likely to be more accurate than those for a calendar year further away from 2000. This is because all three types of annual weights for these children were adjusted to improve the population coverages for June 2000 (as described earlier in the Weighting section).

### **Estimation of Person Characteristics**

For the estimation of the person characteristics in the SPD panel universe using the SPD second longitudinal file, one must first determine which one of the three panel weights (traditional longitudinal panel weight (SPDLNTRD), 1999 quasi-longitudinal panel weight (SPDQLW99), and 2000 quasi-longitudinal panel weight (SPDQLW00)) to be used based on the advantage and disadvantage discussion provided earlier in the Estimation section. Hereinafter, a panel weight (SPDLNTRD or SPDQLW99 or SPDQLW00) chosen for estimating a characteristic or characteristics of a cohort of people (persons) in the SPD panel universe, it will be generically referred to as “*the longitudinal person weight*” for further discussion. Some basic types of longitudinal estimates (using the SPD second longitudinal file) can be constructed using the longitudinal person weights are described below in terms of estimated numbers.

1. The number of people who have ever experienced a characteristic or situation during a given period of time (for example, the number of people who experience unemployment during 1999). To construct such an estimate, sum the weights over all people who possessed the characteristic of interest at some point during the time period of interest.
2. The amount of a characteristic accumulated by people during a given time period (for example, the amount of unemployment compensation received by unemployed people during 1999). To construct such an estimate, compute the product of the weight times the amount of the characteristic and sum this product over all appropriate people.



3. The average number of consecutive months or years of possession of a characteristic (i.e., the spell length for a characteristic.) For example, one could estimate the average spell of unemployment that elapsed before a person found a new job. (*Note that, similar to the SPD first longitudinal file, the SPD second longitudinal file provides the employment data only in terms of week numbers with and without employment in a given year. Thus, for calculation the average unemployment spell length in a time period of interest, the data user needs to match the sample person's record back to the one on the SIPP longitudinal file to determine the number of spells in the time period and/or needs to make some justifiable approximation on the number of unemployment spells within the time period of interest.*) To construct such an estimate, first identify the sample persons possessing the characteristic at some point during the time period of interest. Then, create two sums of these (longitudinal person) weights: Sum 1 is the sum of the products of the weights times the number of months (or years) the spell lasted, and Sum 2 is the sum of the weights only. The average spell length in months (or years) is given by Sum 1 divided by Sum 2. A person who experienced two spells during the time period of interest would be treated as two persons and appear twice in Sum 1 and Sum 2. An alternate method of calculating the average can be found in the section "Standard Error of a Mean or an Aggregate."

*Note that spells extending before or after the time period of interest are cut off (censored) at the boundaries of the time period. If they are used in estimating average spell length, a downward bias will result.*

4. The number of year-to-year changes in the status of a characteristic (i.e., number of transitions) summed over every set of two consecutive years during the time of interest. To construct such estimate, sum the longitudinal person weights each time a change is reported between two consecutive years during the time period of interest. For example, to estimate the number of persons who changed from receiving any public assistance in 1998 to not receiving in 1999 add together the longitudinal person weights of each person who had such a change.
5. Yearly estimates of a characteristic average over a number of consecutive years. For example, we could estimate the yearly average number of food stamp recipients over 1996 through 1999. To construct such an estimate, first form an estimate for each year in the time period of interest by summing up the longitudinal person weights of those possessed the characteristic of interest. Then sum the yearly estimates and divide by the number of years in the time period of interest.

## **ACCURACY OF ESTIMATES**

SPD estimates are based on a sample; they may differ somewhat from the figures that would have been obtained if a complete census had been taken using the same questionnaire, instructions, and enumerators. There are two types of errors possible in an estimate based on a sample survey:

non-sampling and sampling. We are able to provide estimates of the magnitude of SPD sampling error, but this is not true of non-sampling error. The next sections describe sources of SPD non-sampling error, followed by a discussion of sampling error, its estimation, and its use in data analysis.

Note that estimates from this sample for individual states are subject to very high sampling errors and are not recommended. The state codes on the file are primarily of use for linking respondent characteristics with appropriate contextual variables (e.g., state-specific welfare criteria) and for tabulating data by user-defined groupings of states.

### **Non-sampling Errors**

Non-sampling errors can be attributed to many sources, for examples, inability to obtain information about all cases in the sample, difficulties in precisely stating some definitions, differences in the interpretation of questions, inability or unwillingness on the part of the respondents to provide correct information, inability to recall information, and the following errors made. These errors generally include collection such as in recording or coding the data, processing the data, estimating values for missing data, biases resulting from the differing recall periods caused by the rotation pattern used, and under coverage. Quality control and edit procedures were used to reduce errors made by respondents, coders and interviewers.

Under-coverage in SPD results from missed living quarters and missed people within sample households. It is known that under coverage varies with age, race, and gender. Generally, under coverage is larger for males than for females and larger for Blacks than for non-Blacks. Ratio estimation to independent age-race-gender population controls (benchmark estimates) partially corrects for the bias due to survey under-coverage. However, biases exist in the estimates to the extent that people in missed households or missed people in interviewed households have characteristics different from those of interviewed people in the same age-race-gender group. In addition, the independent population controls used have not been adjusted for under-coverage in the decennial census. The Census Bureau has used complex techniques to adjust the weights for non-response. For an explanation of the techniques used, see the "Non-response Adjustment Methods for Demographic Surveys at the U.S. Bureau of the Census," November 1988, Working Paper 8823, by R. Singh and R. Petroni. An example of successfully avoiding bias can be found in "Current Non-response Research for the Survey of Income and Program Participation" (paper by Petroni, presented at the Second International Workshop on Household Survey Non-response, October 1991). The procedure for calculating the longitudinal person weights on the SPD second longitudinal file was derived based on such complex techniques.

Unlike SIPP data that can be analyzed from a cross-sectional or longitudinal view point, the SPD data are solely longitudinal and must be used as such. Thus, the income and poverty estimates in a given single year may not be comparable with those from other surveys such as the Current Population Survey (CPS) and the SIPP. This is principally attributable to the fact that the sample per se and the longitudinal person weights on the SPD second longitudinal file essentially represents just the cohort of people around March 1993. As the SPD sample aged more, it will

become less adequate to represent the more current population (say, the 1998 population). In addition, the high non-response rate (roughly 50 percent) in the SPD may reduce the degree of the effectiveness of the non-interview adjustment process to fully compensate for differential attrition. Note that the non-response rate has three components: 27 percent sample loss inherited from the SIPP, 14 percent occurred from the SPD Bridge interview, and an additional 9 percent occurred at the SPD 1998 interview.

### **Comparability with Other Estimates**

Caution should be exercised when comparing data from this file with data from SIPP publications or with data from other surveys, such as Current Population Survey (CPS). The comparability problems are caused by such sources as the seasonal patterns for many characteristics, different non-sampling errors, and different concepts and procedures. Refer to the *SIPP Quality Profile* for known differences with data from other sources and further discussion.

### **Sampling Variability**

Standard errors indicate the magnitude of the sampling error. They also partially measure the effect of some non-sampling errors in response and enumeration, but do not measure any systematic biases in the data. The standard errors for the most part measure the variations that occurred by chance because a sample rather than the entire population was surveyed.

## **USES AND COMPUTATION OF STANDARD ERRORS**

### **Confidence Intervals**

The sample estimate and its standard error enable one to construct confidence intervals (ranges that would include the average result of all possible samples with a known probability). For example, if all possible samples were selected, each of these being surveyed under essentially the same conditions and

using the same sample design, and if an estimate and its standard error were calculated from each sample, then:

1. Approximately 90 percent of the intervals from 1.645 standard errors below the estimate to 1.645 standard errors above the estimate would include the average result of all possible samples.
2. Approximately 95 percent of the intervals from 1.960 standard errors below the estimate to 1.960 standard errors above the estimate would include the average result of all possible samples.

The average estimate derived from all possible samples is or is not contained in any particular computed interval. However, for a particular sample, one can say with a specified confidence that the average estimate derived from all possible samples is included in the confidence interval.

## **Hypothesis Testing**

Standard errors may also be used for hypothesis testing, a procedure for distinguishing between population characteristics using sample estimates. The most common types of hypotheses tested are the population characteristics are identical versus they are different. Tests may be performed at various levels of significance, where a level of significance is the probability of concluding that the characteristics are different when, in fact, they are identical.

To perform the most common test, compute the difference  $X_A - X_B$ , where  $X_A$  and  $X_B$  are sample estimates of the characteristics of interest. A later section explains how to derive an estimate of the standard error of the difference  $X_A - X_B$ . Let that standard error be  $s_{DIFF}$ . If  $X_A - X_B$  is between  $-1.645$  times  $s_{DIFF}$  and  $+1.645$  times  $s_{DIFF}$ , no conclusion about the characteristics is justified at the 10 percent significance level. If, on the other hand,  $X_A - X_B$  is smaller than  $-1.645$  times  $s_{DIFF}$  or larger than  $+1.645$  times  $s_{DIFF}$ , the observed difference is significant at the 10 percent level. In this event, it is commonly accepted practice to say that the characteristics are different. We recommend that users report only those differences that are significant at the 10 percent level or better. Of course, sometimes this conclusion will be wrong. When the characteristics are, in fact, the same, there is a 10 percent chance of concluding that they are different.

Note that as more tests are performed, more erroneous significant differences will occur. For example, at the 10 percent significance level, if 100 independent hypothesis tests are performed in which there are no real differences, it is likely that about 10 erroneous differences will occur. Therefore, the significance of any single test should be interpreted cautiously.

## **Caution Concerning Small Estimates and Small Differences**

Because of the large standard errors involved, there is little chance that estimates will reveal useful information when computed on a base smaller than 200,000. Also, non-sampling error in one or more of the small number of cases providing the estimate can cause large relative error in that particular estimate. Therefore, care must be taken in the interpretation of small differences since even a small amount of non-sampling error can cause a borderline difference to appear significant or not, thus distorting a seemingly valid hypothesis test.

## **Standard Error Parameters**

Most SPD estimates have greater standard errors than those obtained through a simple random sample because clusters of living quarters are sampled for the SIPP, SPD Bridge, and SPD 1998 to SPD 2000. To derive standard errors that would be applicable to a wide variety of estimates

and could be prepared at a moderate cost, a number of approximations were required. Estimates with similar standard error behavior were grouped together and two parameters (denoted  $a$  and  $b$ ) were developed to approximate the standard error behavior of each group of estimates. Because the actual standard error behavior was not identical for all estimates within a group, the standard errors computed from these parameters provide an indication of the order of magnitude of the standard error for any specific estimate. These  $a$  and  $b$  parameters vary by characteristic and by demographic subgroup to which the estimate applies. The  $a$  and  $b$  parameters are also known as “generalized variance parameters.”

For the SPD second longitudinal file, the  $a$  and  $b$  parameters for various groups of the populations and the six types of weights described below are provided in Tables 3 to 5. Hereinafter, *the  $a$  and  $b$  parameters in Tables 3 to 5* will be referred to as *the base  $a$  and  $b$  parameters*.

As described earlier in the Weighting section, there are six types of weights on the SPD second longitudinal file as listed below.

1. Traditional longitudinal panel weight (SPDLNWGT)
2. Longitudinal annual weight (SPDAWTRD) associated with the traditional longitudinal panel weight
3. 1999 Quasi-longitudinal panel weight (SPDQLW99)
4. Longitudinal annual weight (SPDAWQ99) associated with the 1999 quasi-longitudinal panel weight
5. 2000 Quasi-longitudinal panel weight (SPDQLW00)
6. Longitudinal annual weight (SPDAWQ00) associated with the 2000 quasi-longitudinal panel weight

The applications for each of the six weights are described earlier in the Estimation section and the Longitudinal Annual Weight with Non-original Sample Children section.

The generalized variance parameters (  $a$  and  $b$  ) for the traditional longitudinal panel weight (SPDLNWGT) and the associated longitudinal annual weight (SPDAWTRD) are the same as provided in Table 3. Similarly, the generalized variance parameters (  $a$  and  $b$  ) for the 1999 quasi-longitudinal panel weight (SPDQLW99) and the associated longitudinal annual weight (SPDAWQ99) are the same as provided in Table 4, and the generalized variance parameters (  $a$  and  $b$  ) for the 2000 quasi-longitudinal panel weight (SPDQLW00) and the associated longitudinal annual weight (SPDAWQ00) are the same as provided in Table 5. The sample sizes associated with longitudinal annual weights, SPDAWTRD, SPDAWQ99, and SPDAWQ00 are slightly larger than the sample sizes associated with the longitudinal or quasi-longitudinal panel weights, SPDLNWGT, SPDQLW99, and SPDQLW00, respectively because some of the non-

original sample children were included as part of the longitudinal annual sample as described earlier in the Weighting section. Therefore, the generalized variances parameters provided in Tables 3 to 5 will produce slightly more conservative variance estimates (associated with children aged  $\leq 8$  in 2000) using the longitudinal annual weights (SPDAWTRD, SPDAWQ99, and SPDAWQ00)

### Computation of Standard Error Parameters

In this section we discuss the adjustment of base  $a$  and  $b$  parameters (Tables 3 to 5) to provide  $a$  and  $b$  parameters appropriate for each type of longitudinal described in the section "Estimation of Person Characteristics." Later sections will discuss the use of the adjusted parameters in various formulas to compute standard errors of estimated numbers, percents, averages, etc. Tables 3 to 5 provides the base  $a$  and  $b$  parameters needed to compute the approximate standard errors for estimates.

The creation of appropriate  $a$  and  $b$  parameters for the types of estimates discussed in the section "Estimation of Person Characteristics" is described below. It is assumed that the full sample is used for the estimation.

1. The number of people who have ever experienced a characteristic during a given time period. The appropriate  $a$  and  $b$  parameters are taken directly from Tables 3 to 5 (the base  $a$  and  $b$  parameters). The choice of parameter depends on the characteristic of interest and the demographic subgroup of interest.
2. Amount of a characteristic accumulated by people during a given time period. The appropriate  $a$  and  $b$  parameters are also taken directly from Tables 3 to 5.
3. The average number of consecutive months or years of possession of a characteristic per spell (i.e., the average spell length for a characteristic) during a given time period. Start with the appropriate base  $a$  and  $b$  parameters from Tables 3 to 5. The parameters are then inflated by an additional factor,  $g$  to account for persons who experience multiple spells during the time period of interest. The  $g$  factor is computed by Formula 1 below.

$$g = \frac{\sum_{i=1}^n m_i^2}{\sum_{i=1}^n m_i} \quad (1)$$

where there are  $n$  persons with at least one spell and  $m_i$  is the number of spells experienced by person  $i$  during the time period of interest.

4. The number of year-to-year changes in the status of a characteristic (i.e., number of transitions) summed over every set of two consecutive years during the time period of interest. Obtain a set of adjusted  $a$  and  $b$  parameters exactly as just described in 3, then multiply these parameters by an additional factor of 2.0. The factor of 2.0 is based on the assumption that each spell produces two transitions within the time period of interest.
5. Yearly estimates of characteristic averaged over a number of consecutive years. Appropriate base  $a$  and  $b$  parameters are taken directly from Tables 3 to 5.

### Standard Errors of Estimated Numbers

The approximate standard error  $s_x$  of an estimated number  $x$  of people, families and so forth, can be obtained by using Formula 2 provided below.

$$s_x = \sqrt{ax^2 + bx} \quad (2)$$

Here  $a$  and  $b$  are the standard error parameters associated with the particular type of characteristic for the appropriate longitudinal time period. For the analysis using the SPD data on the SPD second file, the  $a$  and  $b$  parameters are provided in Tables 3 to 5.

An illustration would be to suppose that using the SPD second longitudinal file data, the estimate of the number of people ever receiving Social Security since 1993 and continuing to 1999 is 34,122,000 which was derived using the traditional longitudinal panel weight. The appropriate  $a$  and  $b$  parameters to use in calculating a standard error for the estimate are obtained from Table 3. They are  $a = -0.0000987$ ,  $b = 16,837$ . Using Formula (2), the approximate standard error  $s_x$  is

$$s_x = \sqrt{(-0.0000987)(34,122,000)^2 + (16,837)(34,122,000)} = 677,934 \text{ people}$$

The 90-percent confidence interval as shown by the data is from 33,006,799 to 35,237,201. Therefore, a conclusion that the average estimate derived from all possible samples lies within a range computed in this way would be correct for roughly 90 percent of all samples. Similarly, the 95-percent confidence interval as shown by the data is from 32,793,249 to 35,450,751 and we could conclude that the average estimate derived from all possible samples lies within this interval.

### Standard Error of a Mean or an Aggregate

A mean  $\bar{x}$  is defined here to be the average quantity of some characteristic (other than the number of people, families, or households) per person, family, or household. An aggregate  $k$  is defined to be the total quantity of some characteristic summed over all units in a sub-population.

For example, a mean could be the average annual income of females age 25 to 34. The standard error  $s_{\bar{x}}$  of a mean can be approximated by Formula 3 and the standard error  $s_k$  of an aggregate can be approximated by Formula 4. Because of the approximations used in developing Formulas 3 and 4, an estimate of the standard error of the mean or aggregate obtained from these formulas will generally underestimate the true standard error. The formula used to estimate the standard error  $s_{\bar{x}}$  of a mean  $\bar{x}$  is

$$s_{\bar{x}} = \sqrt{\left(\frac{b}{y}\right) s^2} \quad (3)$$

where  $y$  is the base  $s^2$  is the estimated population variance of the characteristic and  $b$  is the standard error parameter associated with the type of the characteristic. The standard error  $s_k$  of an aggregate  $k$  is

$$s_k = \sqrt{by s^2} \quad (4)$$

The population variance  $s^2$  may be estimated by one of two methods: the first method uses data that has been grouped into intervals, the second method uses ungrouped data. The second method is recommended because it is more precise. However, the first method will be easier to implement if grouped data are already being used as part of the analysis. In both methods, let  $x_i$  denote the value of the characteristic for the  $i^{th}$  person.

To use the first method, the range of values for the characteristic is divided into  $c$  intervals, where the lower and upper boundaries of interval  $j$  are  $Z_{j-1}$  and  $Z_j$ , respectively. Each person is placed into one of the  $c$  groups such that the value of the characteristic,  $x_i$  is between  $Z_{j-1}$  and  $Z_j$ . The estimated population variance,  $s^2$  is then given by Formula 5 below.

$$s^2 = \sum_{j=1}^c p_j m_j^2 - \bar{x}^2 \quad (5)$$

where  $p_j$  is the estimated proportion of people in group  $j$  (based on weighted data), and  $m_j$  is given by the equation below.

$$m_j = \frac{Z_{j-1} + Z_j}{2}, \text{ for } j = 1, 2, \dots, c$$



The most representative value of the characteristic in group  $j$  is assumed to be  $m_j$ . If group  $c$  is open-ended, that is, no upper interval boundary exists, then an approximate value for  $m_c$  is by the equation below.

$$m_c = \left(\frac{3}{2}\right) Z_{c-1}$$

The mean  $\bar{x}$  can be obtained using Formula 6 below.

$$\bar{x} = \sum_{j=1}^c p_j m_j \quad (6)$$

In the second method, the estimated population variance  $s^2$  is given Formula 7 below.

$$s^2 = \frac{\sum_{i=1}^n w_i x_i^2}{\sum_{i=1}^n w_i} - \bar{x}^2 \quad (7)$$

where there are  $n$  sample people with the characteristic of interest and  $w_i$  is the final weight for person  $i$ . The mean  $\bar{x}$  can be obtained from Formula 8 below.

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \quad (8)$$

Note that, by definition,  $y$  (the size of the base) in Formulas 3 and 4 can be obtained from the equation below.

$$y = \sum_{i=1}^n w_i$$

An illustration of Method 1 would be to suppose that, based on the traditional longitudinal panel weight, the 1999 distribution of annual incomes is given in Table 2 for people aged 25 to 34 who were employed for all 12 months of 1999. The mean annual cash income from Formula 8 is

$$\bar{x} = \frac{1,371}{39,851}(2,500) + \frac{1,651}{39,851}(6,250) + \dots + \frac{1,493}{39,851}(105,000) = \$26,717$$

Using Formula 7 and the mean annual cash income of \$26,717 the estimated population variance,  $s^2$  is

$$s^2 = \frac{1,371}{39,851}(2,500)^2 + \frac{1,651}{39,851}(6,250)^2 + \dots + \frac{1,493}{39,851}(105,000)^2 = 468,331,633$$

The appropriate  $b$  parameter from Table 3 is 9,193. Now, using Formula 3, the estimated standard error of the mean is

$$s_{\bar{x}} = \sqrt{\frac{9,193}{39,851,000}(468,331,633)} = \$329$$

An illustration of Method 2 would be to suppose that we are interested in estimating the average length of spell of receiving public assistance during 1992-1994 (just prior to the Welfare Reform) for a given sub-population using the traditional longitudinal panel weight. Also, suppose there are only 10 sample persons in the sub-population who were public assistance recipients. (This example is for illustrative purpose only; in reality, 10 sample units or cases would be too few for a reliable estimate.) The number of consecutive years of receiving public assistance during 1992-1994 are given for each sample person in the table below. (Caveat - In reality, only the total number of months of receiving public assistance in a given year is available in the SPD second longitudinal file. Thus, the actual number of spells in a time period of interest is not known or equivalently the actual spell length is not known. Consequently, to use such data in the SPD second longitudinal file for assessing average spell length in a time period of interest, it is the responsibility of the data user to match back the sample person record to the one in the SIPP longitudinal file to determine the number of spells in the time period of interest and/or make his/her own justifiable assumption on what should be the number of spells in the time period of interest given the total number of months in a year that a sample person possessed a spell characteristic, e.g., receiving public assistance.)

Sample Person Number	Number of Spells During 1992-1994	Spell Lengths in Months	Final Longitudinal Weight
1	2	12, 6	5300
2	1	2	7100
3	1	5	4900
4	2	3, 6	6500
5	1	13	4700
6	1	14	5500
7	2	3, 6	4100
8	1	24	4200
9	1	6	4500
10	1	4	6100

Using Formula 8, the average spell  $\bar{x}$  of receiving public assistance is estimated to be

$$\bar{x} = \frac{5300 \times 12 + 5300 \times 6 + 7100 \times 2 + \dots + 6100 \times 4}{5300 + 5300 + 7100 + \dots + 6100} = \frac{472800}{68800} = 6.872 \text{ months}$$

The standard error  $s_{\bar{x}}$  will be computed by Formula 3. First, estimate the population variance  $s^2$  by Formula 7

$$s^2 = \frac{5300 \times 12^2 + 5300 \times 6^2 + 7100 \times 2^2 + \dots + 6100 \times 4^2}{5300 + 5300 + 7100 + \dots + 6100} - 6.872^2 = 41.92 \text{ months}^2$$

Next, the base  $b$  parameter from Table 3 is 17,740. To account for the multiple number of spells during 1992-1994 of three sample persons (two spells for Sample Persons 1, 4, and 7), multiply the base  $b$  parameter by a factor  $g$  computed from Formula 1 as shown below.

$$g = \frac{2^2 + 1 + 1 + 2^2 + 1 + 1 + 2^2 + 1 + 1 + 1}{2 + 1 + 1 + 2 + 1 + 1 + 2 + 1 + 1 + 1} = 1.462$$

Therefore, the adjusted b parameter is  $17740 \times 1.462 = 25936$  and the standard error  $s_{\bar{x}}$  of the mean is

$$s_{\bar{x}} = \sqrt{\frac{25936}{68800} \times 41.92} = 3.975 \text{ months}$$

### Standard Errors of Estimated Percentages

This section refers to the percentages of a group of people, families, or households possessing a particular attribute and to percentages of money or related concepts. The reliability of an estimated percentage, computed using sample data for both numerator and denominator, depends upon both the size of the percentage and the size of the total upon which the percentage is based. Estimated percentages are relatively more reliable than the corresponding estimates of the numerators of the percentages, particularly if the percentages are more than 50 percent. For example, the percent estimate of employed people is more reliable than the estimated number of employed people. When the numerator and denominator of the percentage have different parameters, use the parameter of the numerator. If proportions are presented instead of percentages, note that the standard error of a proportion is equal to the standard error of the corresponding percentage divided by 100.

There are two types of percentages commonly estimated. The first type is the percentage of people sharing a particular characteristic such as the percentage of people owning their own home or the percentage of 1998 food stamp recipients who were also receiving food stamps in 1999. The second type is the percentage of money or some similar concept held by a particular group of people or held in a particular form. Examples are the percentage of wealth held by people with high income and the percentage of annual income received by females.

For the percentage of people, the approximate standard error,  $s_{x,p}$ , of the estimated percentage,  $p$ , can be obtained by Formula 9 below.

$$s_{x,p} = \sqrt{\frac{b}{x} p(100 - p)} \quad (9)$$

Here,  $x$  is the base of the percentage  $p$  is the percentage ( $0 < p < 100$ ), and  $b$  is parameter for the numerator of the percentage calculation. For the analysis using the SPD data and the SPD second longitudinal file, the  $b$  parameters are provided in Tables 3 to 5.

An illustration would be to suppose that, based on the traditional longitudinal panel weight, in 1999, an estimate of number of male aged 22 to 55 was 46,023,000. Among all the males in this age group, an estimate of 2.4 percent was unemployed. The  $b$  parameter associated with the numerator (the number of unemployed male) is 9,193 (from Table 3). Using Formula 9, the approximate standard error  $s_{x,p}$  is

$$s_{x,p} = \sqrt{\frac{9,193}{46,023,000} (2.4)(1 - 2.4)} = 0.22\%$$

Consequently, the 90-percent confidence interval for the unemployment estimate is 2.0% to 2.8%.

To calculate the percentages of money, the formula is more complicated. A percentage of money will usually be estimated in one of two ways. It may be the ratio,  $p_M$  of two aggregates as defined in Formula 10 below.

$$p_M = 100 \left( \frac{X_A}{X_N} \right) \quad (10)$$

or it may be the ratio,  $p_M$  of two means with an adjustment,  $\hat{p}_A$  for different bases as defined in Formula 11 below.

$$p_M = 100 \left( \frac{\bar{X}_A}{\bar{X}_N} \right) \hat{p}_A \quad (11)$$

where  $X_A$  and  $X_N$  in Formula 10 are aggregate money figures,  $\bar{X}_A$  and  $\bar{X}_N$  in Formula 11 are mean money figures, and  $\hat{p}_A$  is the estimated number in Group A divided by the estimated number in Group

N. In either way of estimating  $p_M$  (Formula 10 or 11), we estimate the standard error  $s_{p_M}$  of  $p_M$  using Formula 12 provided below.

$$s_{p_M} = \sqrt{\left(\frac{\hat{p}_A \bar{X}_A}{\bar{X}_N}\right)^2 \left[ \left(\frac{s_{\hat{p}_A}}{\hat{p}_A}\right)^2 + \left(\frac{s_{\bar{X}_A}}{\bar{X}_A}\right)^2 + \left(\frac{s_{\bar{X}_N}}{\bar{X}_N}\right)^2 \right]} \quad (12)$$

where  $s_{\hat{p}_A}$  is the standard error of  $\hat{p}_A$ ,  $s_{\bar{X}_A}$  is the standard error of  $\bar{X}_A$  and  $s_{\bar{X}_N}$  is the standard error of  $\bar{X}_N$ . To calculate  $s_{\hat{p}_A}$ , use Formula 9. The standard errors  $s_{\bar{X}_A}$  and  $s_{\bar{X}_N}$  are calculated using Formula 3.

Note that there is frequently some correlation among the characteristics estimated by  $\hat{p}_A$ ,  $\bar{X}_A$ , and  $\bar{X}_N$ . These correlations, if present, will cause a tendency toward overestimates or underestimates, depending on the relative sizes of the correlations and whether they are positive or negative.

An illustration would be to suppose that, based on the traditional longitudinal weight, in 2000, an estimated 8.8% of males aged 16 and over was Black, the mean annual earning of these Black males was \$15,456, the mean annual earning of all males aged 16 and over was \$22,932, and the corresponding standard errors are 0.37 percent, \$432, and \$324, respectively. Then, the percent ( $p_M$ ) of male earnings made by Blacks in 1998 per Formula 11 is

$$p_M = 100 \left( \frac{15,456}{22,932} \right) (0.088) = 5.9\%$$

Using Formula 12, the approximate standard error,  $s_{p_M}$  is

$$s_{p_M} = \sqrt{\left(\frac{(0.088)(15,456)}{22,932}\right)^2 \left[ \left(\frac{0.0037}{0.088}\right)^2 + \left(\frac{432}{15,456}\right)^2 + \left(\frac{324}{22,932}\right)^2 \right]} = 0.31\%$$

## Standard Error of a Difference

The standard error  $s_{x-y}$  of a difference between two sample estimates  $x$  and  $y$  is equal to

$$s_{x-y} = \sqrt{s_x^2 + s_y^2 - 2rs_x s_y} \quad (13)$$

where  $s_x$  and  $s_y$  are the standard errors of the estimates  $x$  and  $y$ . The estimates can be numbers, averages, percents, ratios, etc. The correlation between  $x$  and  $y$  is represented by  $r$  ( $0 \leq r \leq 1$ ). If  $r$  is assumed to be zero and the true correlation is really positive (negative), then this assumption will result in a tendency toward overestimates (underestimates) of the true standard error.

An illustration would be to suppose that we are interested in the difference in the average annual number of adult males (aged 16 and above) versus adult females with annual cash income above \$9,000 in 1998. An estimate of the number of adult people in this income bracket has been obtained for both males and females. For females, the estimate is 1,619,000. A similar estimate for males is 2,198,000. The difference in estimates is 579,000.

The standard error of the adult female estimate is computed next. The  $a$  and  $b$  parameters from Table 3 for females are -0.0001027 and 9,193, respectively. Based on Formula 2, the standard error,  $s_x$  of the female estimate is

$$s_x = \sqrt{(-0.0001027)(1,619,000)^2 + (9,193)(1,619,000)} = 120,890$$

Similarly, the  $a$  and  $b$  parameters from Table 3 for males are -0.0001137 and 9,193, respectively. Based on Formula 2, the standard error,  $s_y$  of the male estimate is

$$s_y = \sqrt{(-0.0001137)(2,198,000)^2 + (9,193)(2,198,000)} = 140,203$$

Now, the standard error of the difference is computed using the above two standard errors. The correlation  $r$  for this example is assumed to be zero. The standard error,  $s_{x-y}$  of the difference is computed by Formula 13 as shown below.

$$s_{x-y} = \sqrt{(120,890)^2 + (140,203)^2} = 185,125$$

Suppose that it is desired to test at the 10 percent significance level whether the number of adult males and females with monthly cash income above \$9,000 were different in 1998, one can compare the difference of 579,000 to the product  $1.645 \times 185,125 = 304,531$ . Since the difference is larger than 1.645 times the standard error ( $s_{x-y}$ ) of the difference, the data allow us to conclude that, in 1998, the number of adult males with annual cash income above \$90,000 is significantly higher than the number of the adult females at the 10 percent confidence level.

## Standard Error of a Median

The median quantity,  $X_{med}$  of some item (characteristic),  $X$  such as income for a given group of people, families, or households is that quantity such that at least half the group has as much or more and at least half the group has as much or less. The sampling variability of an estimated median  $\hat{X}_{med}$  depends upon the form of the distribution of the item as well as the size of the group. To estimate the median ( $X_{med}$ ) and the standard error of the median  $s_{X_{med}}$  the procedure described below may be used.

The median ( $X_{med}$ ) like the mean, can be estimated using either data which has been grouped into intervals (e.g., income intervals) or ungrouped data. If grouped data are used, the median ( $X_{med}$ ) is estimated using either Formula 15 or 16 with  $p = 0.5$ . If ungrouped data are used, the data records are ordered based on the value of the item (e.g., income level), then the estimated median is the value of the item such that the weighted estimate of 50 percent of the sub-population falls at or below that value and 50 percent is at or above that value. The method of standard error computation presented here requires the use of grouped data, because it is deemed easier to compute the median by grouping the data and then using Formula 15 or 16.

An approximate method for measuring the reliability of an estimated median ( $\hat{X}_{med}$ ) is to determine a confidence interval about it. (See the section on "Confidence Intervals.") The following procedure (four steps) may be used to estimate the 68-percent confidence limits (i.e., approximately  $\pm$  one standard error from the median) and hence the standard error (of a median) based on sample data.

Step 1 - Determine, using Formula 9, the standard error ( $s_{x,p=50}$ ) of an estimate of 50 percent of the group (sub-population).

Step 2 - Subtract from and add to 50 percent the standard error determined in Step 1 to obtain the percentages associated with the lower and upper limits of the 68-percent confidence interval of the item. Namely, the smaller percentage is  $50 - s_{x,p=50}$  percent, and the larger percentage is  $50 + s_{x,p=50}$  percent.

Step 3 - Using the distribution of the item within the group, calculate the quantity,  $X_{UCL}$  of the item such that the percent of the group owning more of the item is equal to the smaller percentage ( $50 - s_{x,p=50}$ ) found in Step 2. This quantity ( $X_{UCL}$ ) will be the upper limit for the 68-percent confidence interval (assuming that the interval with higher item value is ranked at lower percentile as illustrated in Table 2.) In a similar fashion, calculate the quantity,  $X_{LCL}$  of the item such that the percent of the group owning more of the item is equal to the larger percentage ( $50 + s_{x,p=50}$ ) found in Step 2. This quantity ( $X_{LCL}$ ) will be the lower limit for the 68-percent confidence interval. (Note that a median computed from ungrouped data may or may not fall in this confidence interval).



Step 4 - Divide the difference between the two quantities ( $X_{UCL}$  and  $X_{LCL}$ ) determined in Step 3 by two to obtain the standard error estimate ( $s_{\hat{X}_{med}}$ ) of the median estimate ( $\hat{X}_{med}$ ). Namely,

$$\hat{s}_{\hat{X}_{med}} = \frac{X_{UCL} - X_{LCL}}{2} \quad (14)$$

To perform Step 3, it will be necessary to interpolate, which may be done using different methods. The most common is simple linear interpolation (Formula 15) and Pareto interpolation (Formula 16). The appropriateness of the method depends on the form of the distribution around the median. We recommend Pareto interpolation in most instances. Interpolation is used as follows. The quantity of the item,  $X_{pN}$  such that  $p$  percent own more of the item is

$$X_{pN} = A_1 \exp \left[ \frac{\ln \left( \frac{pN}{N_1} \right)}{\ln \left( \frac{N_2}{N_1} \right)} \ln \left( \frac{A_2}{A_1} \right) \right] \quad (15)$$

if Pareto Interpolation is indicated and

$$X_{pN} = \left[ \frac{pN - N_1}{N_2 - N_1} (A_2 - A_1) + A_1 \right] \quad (16)$$

if linear interpolation is indicated, where  $N$  is the size of the group;  $A_1$  and  $A_2$  are the lower and upper bounds, respectively, of the interval in which  $X_{pN}$  falls;  $N_1$  and  $N_2$  are the estimated numbers of group members owning more than  $A_1$  and  $A_2$ , respectively;  $exp$  refers to the exponential function; and  $Ln$  refers to the natural logarithm function. One should note that a mathematically equivalent result is obtained by using common logarithms (base 10) and antilogarithms.

An illustration would be in order to calculate the standard error of a median, we return to the first example used to illustrate the standard error of a mean. As indicated in Table 2, the size ( $N$ ) of the group is 39,851,000 and the median annual income estimate ( $\hat{X}_{med}$ ) for the group falls in between \$17,500 and \$19,999. With  $p = 0.5$ ,  $A_1 = \$17,500$ ,  $A_2 = \$19,999$ ;  $N_1 = 5,799,000 + 4,730,000 + \dots + 1,493,000 = 22,106,000$ , and  $N_2 = 4,730,000 + 3,723,000 + \dots + 1,493 = 16,307,000$ ; the median annual income estimate,  $\hat{X}_{med}$  for this group is computed using Formula 16 (with  $p = 0.5$ ) to be \$18,316. The standard error estimate ( $s_{\hat{X}_{med}}$ ) of the median annual income

estimate is calculated using the above four step procedure as follows.

Step 1 - Using Formula 9 and the appropriate  $b$  parameter of 9,193, the standard error estimate of 50 percent on a base of 39,851,000 is about 0.8 percentage points, (i.e.,  $s_{x,p=50} = 0.8\%$ ).

Step 2 - Obtain the two percentages associated with the lower and upper limits of the 68-percent confidence: the smaller percentage =  $50 - s_{x,p=50} = 49.2$  and the larger percentage =  $50 + s_{x,p=50} = 50.8$ .

Step 3 - By examining Table 2, we see that the percentage 49.2 falls in the income interval from \$17,500 to \$19,999. Thus as determined previously,  $A_1 = \$17,500$ ,  $A_2 = \$19,999$ ,  $N_1 = 22,106,000$ ,  $N_2 = 16,307,000$ , and  $N = 39,851,000$  and  $p = 49.2$ . Based on Formula 15, the upper bound ( $X_{UCL}$ ) of a 68-percent confidence interval for the median estimate ( $\hat{X}_{med}$ ) is

$$X_{UCL} = 17,500 \exp \left[ \frac{\ln \left( \frac{0.492 \times 39,851,000}{22,106,000} \right)}{\ln \left( \frac{16,307,000}{22,106,000} \right)} \ln \left( \frac{19,999}{17,500} \right) \right] = \$18,446$$

Also by examining Table 2, the 50.8 percent fall in the same income interval. Thus,  $A_1$ ,  $A_2$ ,  $N_1$ , and  $N_2$  are the same as above, but  $p = 0.508$ . The lower bound ( $X_{LCL}$ ) of a 68-percent confidence interval for the median ( $\hat{X}_{med}$ ) is

$$X_{LCL} = 17,500 \exp \left[ \frac{\ln \left( \frac{0.508 \times 39,851,000}{22,106,000} \right)}{\ln \left( \frac{16,307,000}{22,106,000} \right)} \ln \left( \frac{19,999}{17,500} \right) \right] = \$18,189$$

Step 4 - Based on Formula 14, the standard error estimate ( $s_{\hat{X}_{med}}$ ) of the median annual income estimate ( $\hat{X}_{med}$ ) is

$$s_{\hat{X}_{med}} = \frac{\$18,446 - \$18,189}{2} = \$129$$

If the linear interpolation is used, the median is then estimated using Formula 16 to be \$18,440 and the 68-percent confidence interval of the estimated median is from \$18,302 to \$18,577. The standard error estimate is \$138.

### Standard Error of Ratio of Means or Medians

The standard error for a ratio of means or medians is approximated by Formula 17 provided below.

$$s_{\frac{X}{Y}} = \sqrt{\left(\frac{X}{Y}\right)^2 \left[ \left(\frac{s_X}{X}\right)^2 + \left(\frac{s_Y}{Y}\right)^2 \right]} \quad (17)$$

where  $X$  and  $Y$  are the means or medians, and  $s_X$  and  $s_Y$  are their associated standard errors. Formula 17 assumes that the means or medians are not correlated. If the correlation between the population means or medians estimated by  $X$  and  $Y$  are actually positive (negative), then this procedure will tend to produce overestimates (underestimates) of the true standard error for the ratio of means or medians.

Table 1a - Reference months for each interview month of the SIPP 1992 Panel, SIPP 1993 Panel, SPD Bridge (1997), and SPD 1998 Surveys.

Survey	Months of Interview	Reference Months
SIPP Panel 1992	February 1992 - April 1995	October 1991 - March 1995
SIPP Panel 1993	February 1993 - January 1996	October 1992 - December 1995
SPD Bridge (1997)	April 1997 - June 1997	January 1996 - December 1996 (also January 1995 - December 1995 for SIPP Panel 1992 for only selected questions)
SPD 1998	May 1998 - July 1998	January 1997 - December 1997
SPD 1999	May 1999 - July 1999	January 1998 - December 1998
SPD 2000	May 2000 - July 2000	January 1999 - December 1999





Table 2 - Distribution of annual income among people 25 to 34 years old.

Number of People in Annual Income Interval														
Total Number of People	Under \$5000	\$5000 to \$7499	7500 to \$9999	\$10000 to \$12499	\$12500 to \$14999	\$15000 to \$17499	\$17500 to \$19999	\$20000 to \$29999	\$30000 to \$39999	\$40000 to \$49999	\$50000 to \$59999	\$60000 to \$69999	\$70000 and Over	
	1371	1651	2259	2734	3452	6278	5799	4730	3723	2591	2619	1223	1493	
Number of People (in Thousands)	39851													
Percent with at Least as Much as Lower Bound of Interval	N/A	100.0	96.6	92.4	86.7	79.9	71.2	55.5	40.9	29.1	19.7	13.4	6.8	3.7

Note: This table contains a fictitious distribution of annual income and is used only to illustrate standard error calculation.

Table 3 - SPD Generalize variance parameters for estimates using the traditional longitudinal panel weights (SPDLNWGT) or the longitudinal annual weights (SPDAWTRD) associated with the traditional longitudinal panel weights (SPDLNWGT) on the SPD second longitudinal file.

Characteristic	Parameters	
	a	b
TOTAL OR WHITE PEOPLE		
16+ Program Participation and Benefits, Poverty (3)*		
Both Sexes	-0.0001042	17,740
Male	-0.0002193	17,740
Female	-0.0001984	17,740
16+ Income and Labor Force (5)*		
Both Sexes	-0.0000538	9,193
Male	-0.0001137	9,193
Female	-0.0001027	9,193
16+ Pension Plan** (4)*		
Both Sexes	-0.0000987	16,837
Male	-0.0002083	16,837
Female	-0.0001882	16,837
All Others*** (6)*		
Children Aged Less Than 18		
Both Sexes	-0.0000970	22,354
Male	-0.0002004	22,354
Female	-0.0001878	22,354
Adults Aged 18 and Over		
Both Sexes	-0.0001449	33,436
Male	-0.0002996	33,436



Characteristic	Parameters	
	a	b
Female	-0.0002810	33,436
BLACK PEOPLE		
Poverty (1)*		
Both Sexes	-0.0005483	15,130
Male	-0.0011785	15,130
Female	-0.0010258	15,130
All Others*** (2)*		
Children Aged Less Than 18		
Both Sexes	-0.0003000	8,269
Male	-0.0006441	8,269
Female	-0.0005605	8,269
Adults Aged 18 and Over		
Both Sexes	-0.0004487	12,369
Male	-0.0009634	12,369
Female	-0.0008385	12,369
HOUSEHOLDS		
Total or Whites	-0.0001281	11,363
Black	-0.0007826	7,850

\* For cross-tabulations, use the *a* and *b* parameters of the characteristic with the smaller number within the parentheses.

\*\* Use the “16+ Pension Plan” parameters for pension plan tabulations of people aged 16+ in the labor force. Use the “All Others” parameters for retirement tabulations, 0+ program participation, 0+ benefits, 0+ income, and 0+ labor force tabulations, in addition to any other types of tabulations not specifically covered by another characteristic in this table.

\*\*\* Use the “All Others” parameters for any type of tabulation not specifically covered by another characteristic in this table.

Table 4 - SPD Generalize variance parameters for estimates using the 1999 quasi-longitudinal panel weights (SPDQLW99) or the longitudinal annual weights (SPDAWQ99) associated with the 1999 quasi-longitudinal panel weights (SPDQLW99) on the SPD second longitudinal file.

Characteristic	Parameters	
	a	b
TOTAL OR WHITE PEOPLE		
16+ Program Participation and Benefits, Poverty (3)*		
Both Sexes	-0.0000878	14,937
Male	-0.0001847	14,937
Female	-0.0001671	14,937
16+ Income and Labor Force (5)*		
Both Sexes	-0.0000453	7,740
Male	-0.0000958	7,740
Female	-0.0000864	7,740
16+ Pension Plan** (4)*		
Both Sexes	-0.0000831	14,177
Male	-0.0001753	14,177
Female	-0.0001585	14,177
All Others*** (6)*		
Children Aged Less Than 18		
Both Sexes	-0.0000816	18,821
Male	-0.0001687	18,821
Female	-0.0001582	18,821
Adults Aged 18 and Over		
Both Sexes	-0.0001220	28,152
Male	-0.0002523	28,152

Characteristic	Parameters	
	a	b
Female	-0.0002366	28,152
BLACK PEOPLE		
Poverty (1)*		
Both Sexes	-0.0004617	12,739
Male	-0.0009923	12,739
Female	-0.0008637	12,739
All Others*** (2)*		
Children Aged Less Than 18		
Both Sexes	-0.0002526	6,963
Male	-0.0005423	6,963
Female	-0.0004719	6,963
Adults Aged 18 and Over		
Both Sexes	-0.0003778	10,414
Male	-0.0008111	10,414
Female	-0.0007060	10,414
HOUSEHOLDS		
Total or Whites	-0.0001078	9,567
Black	-0.0006589	6,610

\* For cross-tabulations, use the *a* and *b* parameters of the characteristic with the smaller number within the parentheses.

\*\* Use the “16+ Pension Plan” parameters for pension plan tabulations of people aged 16+ in the labor force. Use the “All Others” parameters for retirement tabulations, 0+ program participation, 0+ benefits, 0+ income, and 0+ labor force tabulations, in addition to any other types of tabulations not specifically covered by another characteristic in this table.

\*\*\* Use the “All Others” parameters for any type of tabulation not specifically covered by another characteristic in this table.

Table 5 - SPD Generalize variance parameters for estimates using the 2000 quasi-longitudinal panel weights (SPDQLW00) or the longitudinal annual weights (SPDAWQ00) associated with the 2000 quasi-longitudinal panel weights (SPDQLW00) on the SPD second longitudinal file.

Characteristic	Parameters	
	a	b
TOTAL OR WHITE PEOPLE		
16+ Program Participation and Benefits, Poverty (3)*		
Both Sexes	-0.0000905	15,404
Male	-0.0001904	15,404
Female	-0.0001723	15,404
16+ Income and Labor Force (5)*		
Both Sexes	-0.0000467	7,982
Male	-0.0000987	7,982
Female	-0.0000891	7,982
16+ Pension Plan ** (4)*		
Both Sexes	-0.0000857	14,620
Male	-0.0001808	14,620
Female	-0.0001634	14,620
All Others *** (6)*		
Children Aged Less Than 18		
Both Sexes	-0.0000842	19,410
Male	-0.0001740	19,410
Female	-0.0001631	19,410
Adults Aged 18 and Over		
Both Sexes	-0.0001259	29,033
Male	-0.0002602	29,033

Characteristic	Parameters	
	a	b
Female	-0.0002440	29,033
BLACK PEOPLE		
Poverty (1)*		
Both Sexes	-0.0004761	13,138
Male	-0.0010234	13,138
Female	-0.0008907	13,138
All Others*** (2)*		
Children Aged Less Than 18		
Both Sexes	-0.0002605	7,180
Male	-0.0005593	7,180
Female	-0.0004867	7,180
Adults Aged 18 and Over		
Both Sexes	-0.0003896	10,740
Male	-0.0008365	10,740
Female	-0.0007281	10,740
HOUSEHOLDS		
Total or Whites	-0.0001112	9,866
Black	-0.0006795	6,816

\* For cross-tabulations, use the *a* and *b* parameters of the characteristic with the smaller number within the parentheses.

\*\* Use the “16+ Pension Plan” parameters for pension plan tabulations of people aged 16+ in the labor force. Use the “All Others” parameters for retirement tabulations, 0+ program participation, 0+ benefits, 0+ income, and 0+ labor force tabulations, in addition to any other types of tabulations not specifically covered by another characteristic in this table.

\*\*\* Use the “All Others” parameters for any type of tabulation not specifically covered by another characteristic in this table.